

개인화된 광고 채널 선택을 위한 데이터 분석 및 딥러닝 모델 제안

이수현

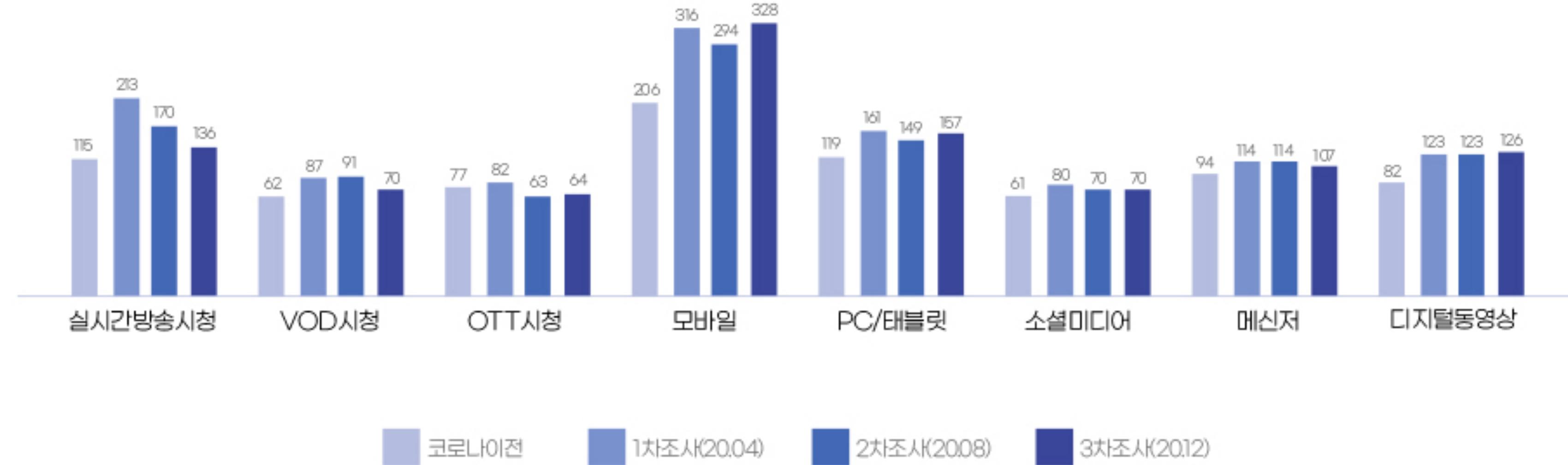
Part0.

Personalized Channel Selection

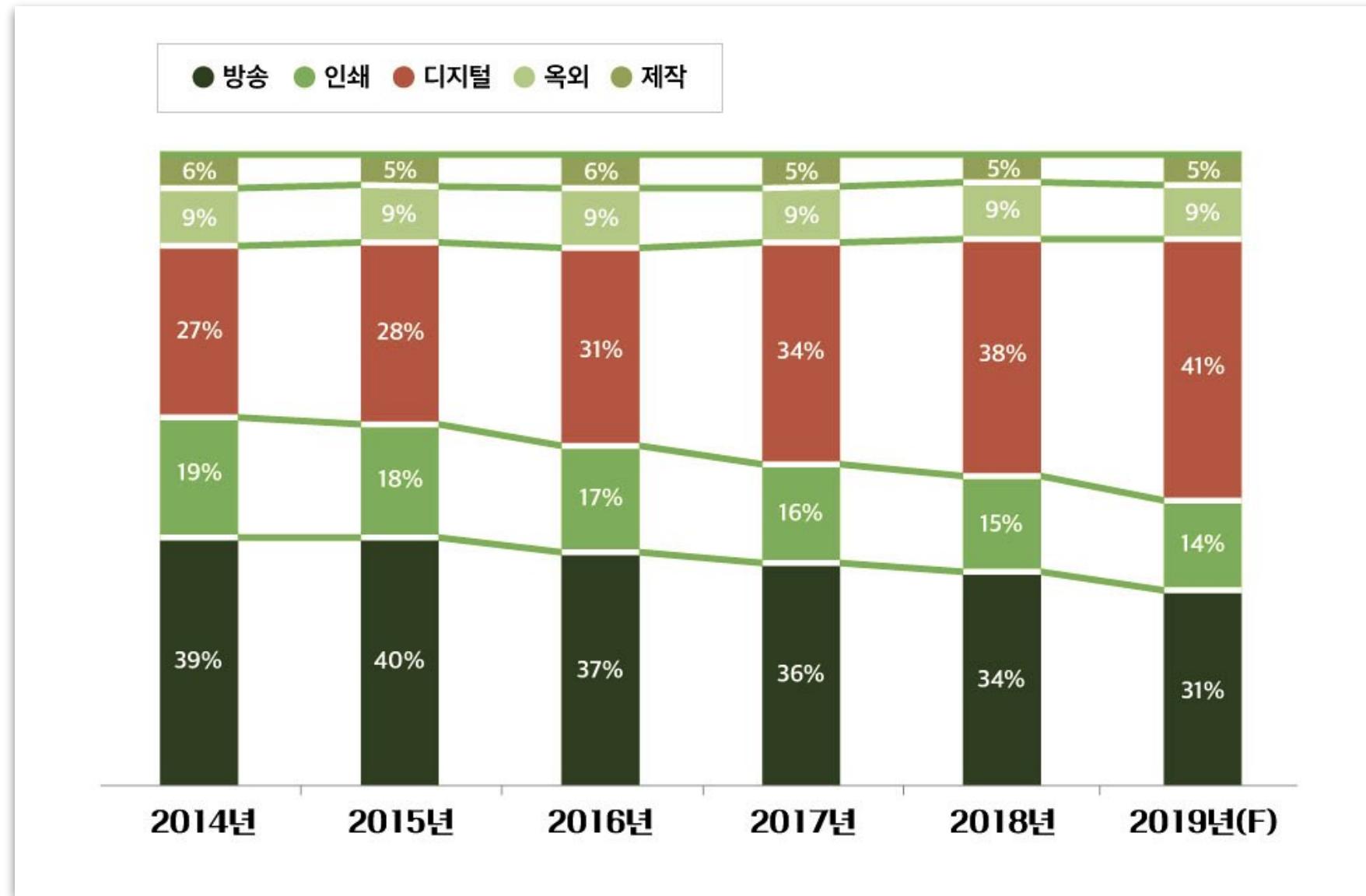
우리가 손에 쥐고 놓지 않는 것 한가지,

스마트폰

최근 1개월 기준 미디어 평균 이용 시간

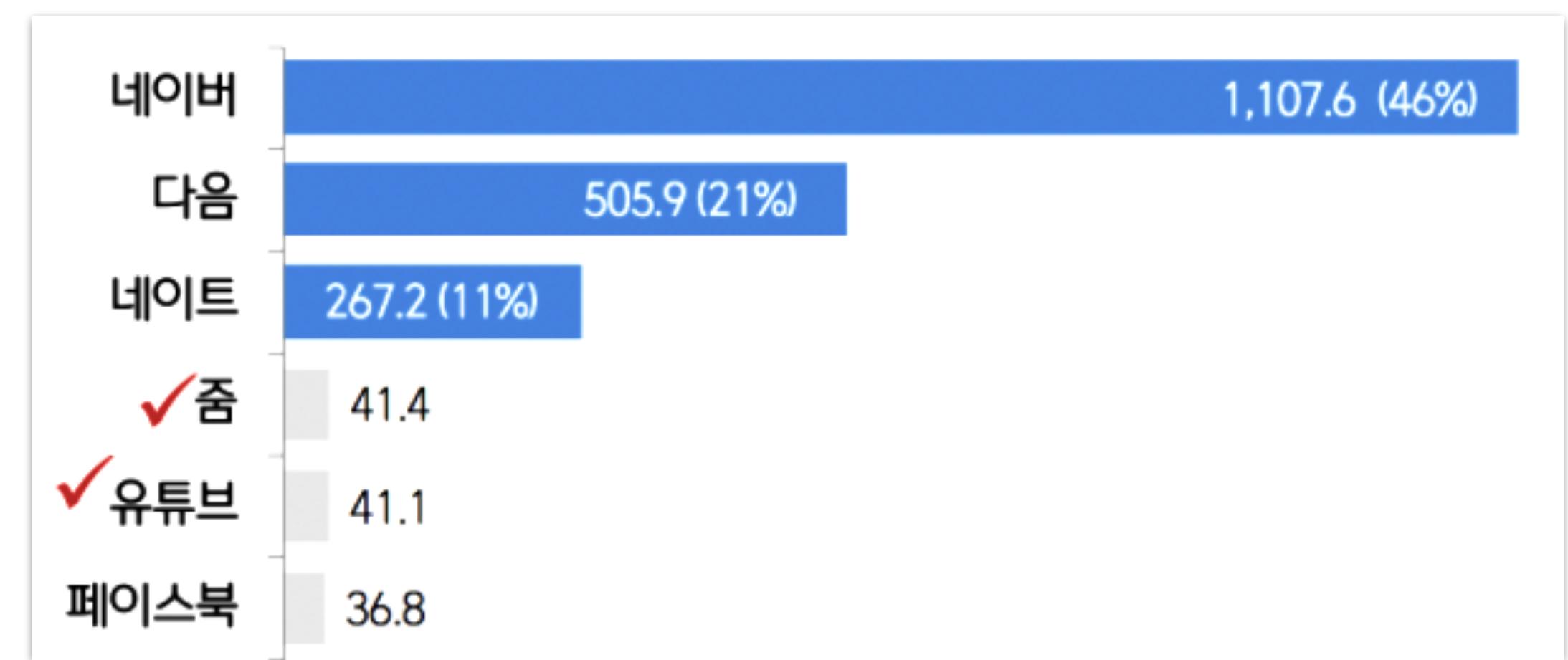


2019 제일기획 매체별 광고비 비율



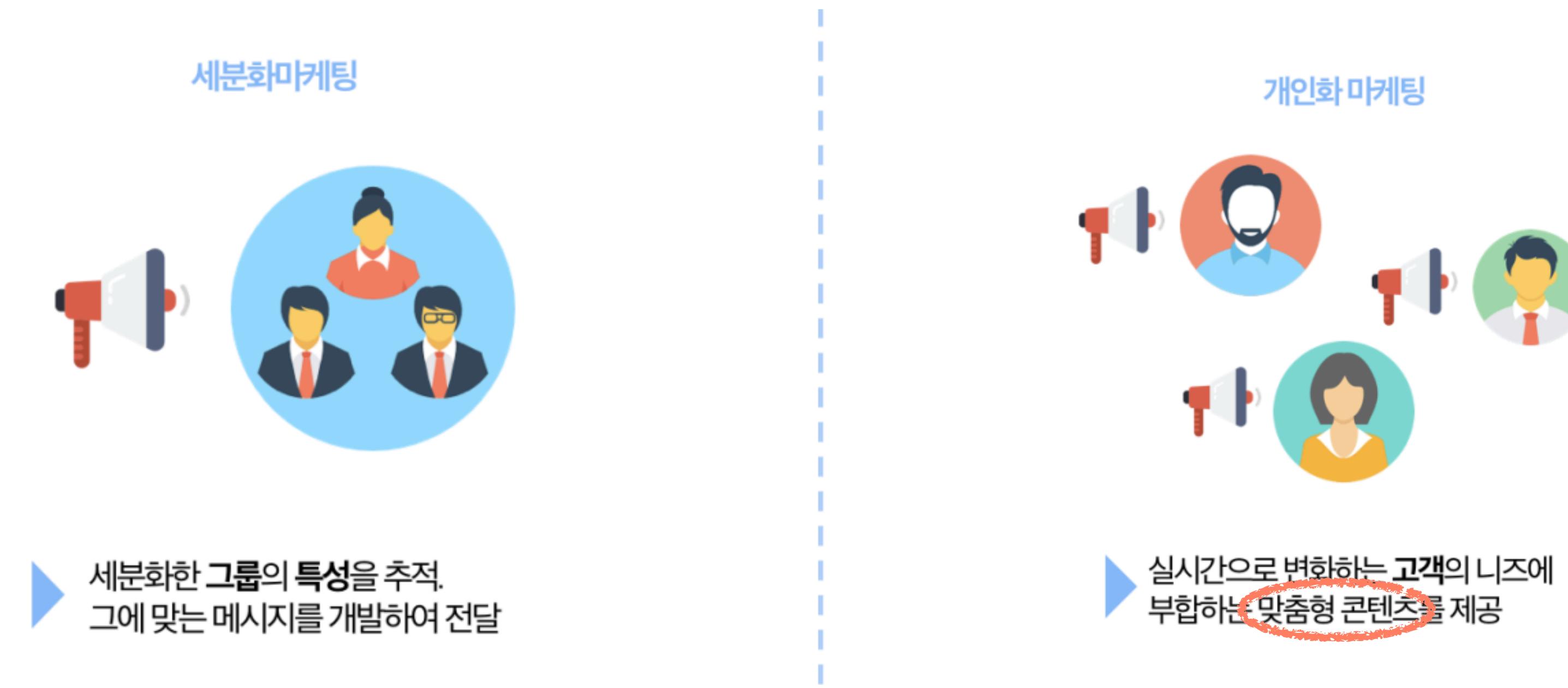
모바일 광고는 가파르게 성장하고 있으며,

2017 Top6 매체별 광고비



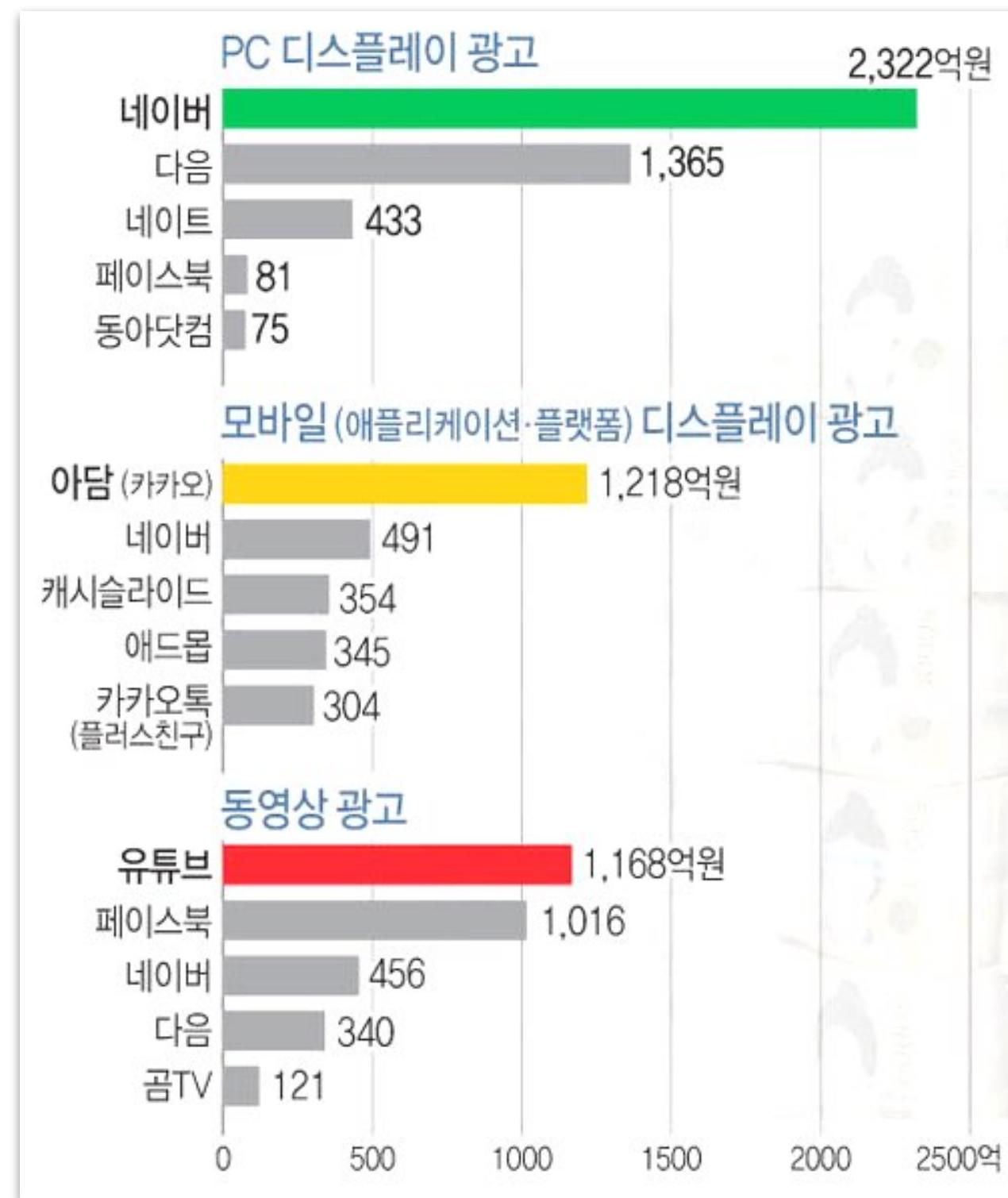
매체별 광고비도 점점 비싸지는 중

비싼 광고비, 보다 효과적으로 쓸 수 있지 않을까?



콘텐츠를 맞춤화 하는 것도 방법이지만,
채널 자체에서부터 광고비를 아낄 수 있다면?

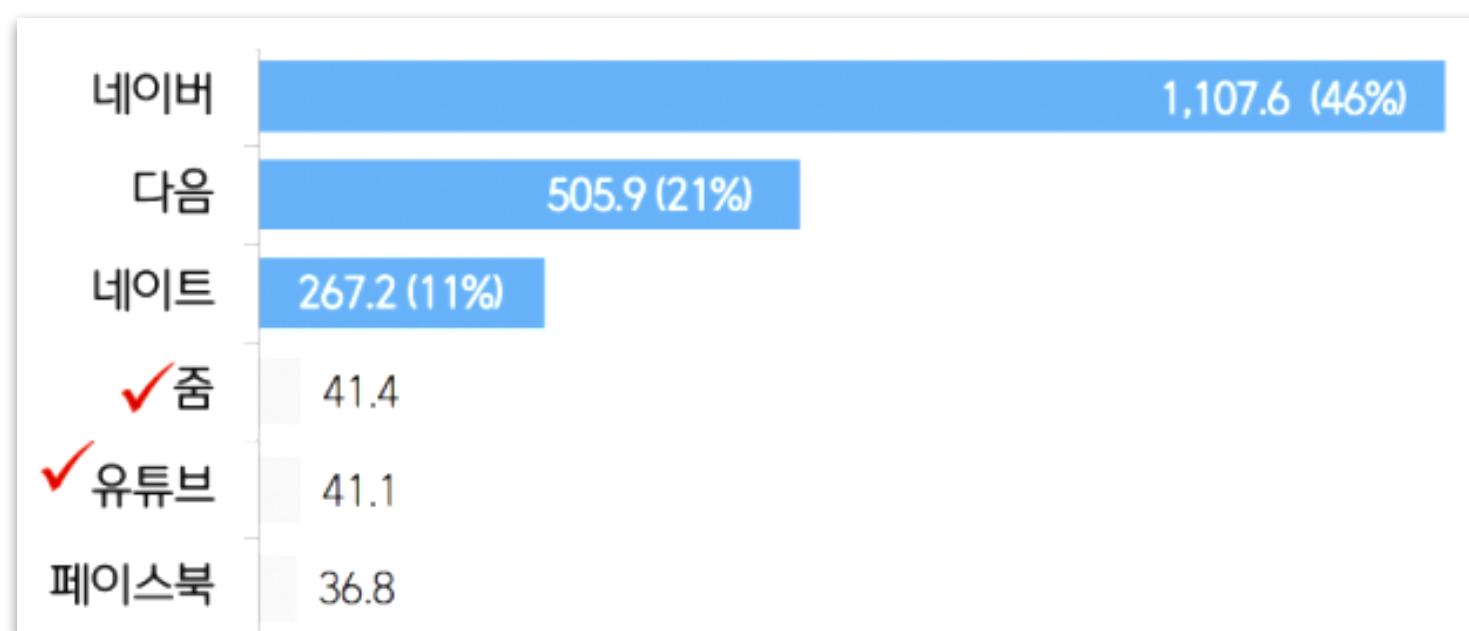
항목별 광고비 지출 상위 5개 매체



비싼 곳에 지출도 가장 많이 하는데,
채널도 개인화 해서
채널비도 효과적으로 써 보는게 어때요?

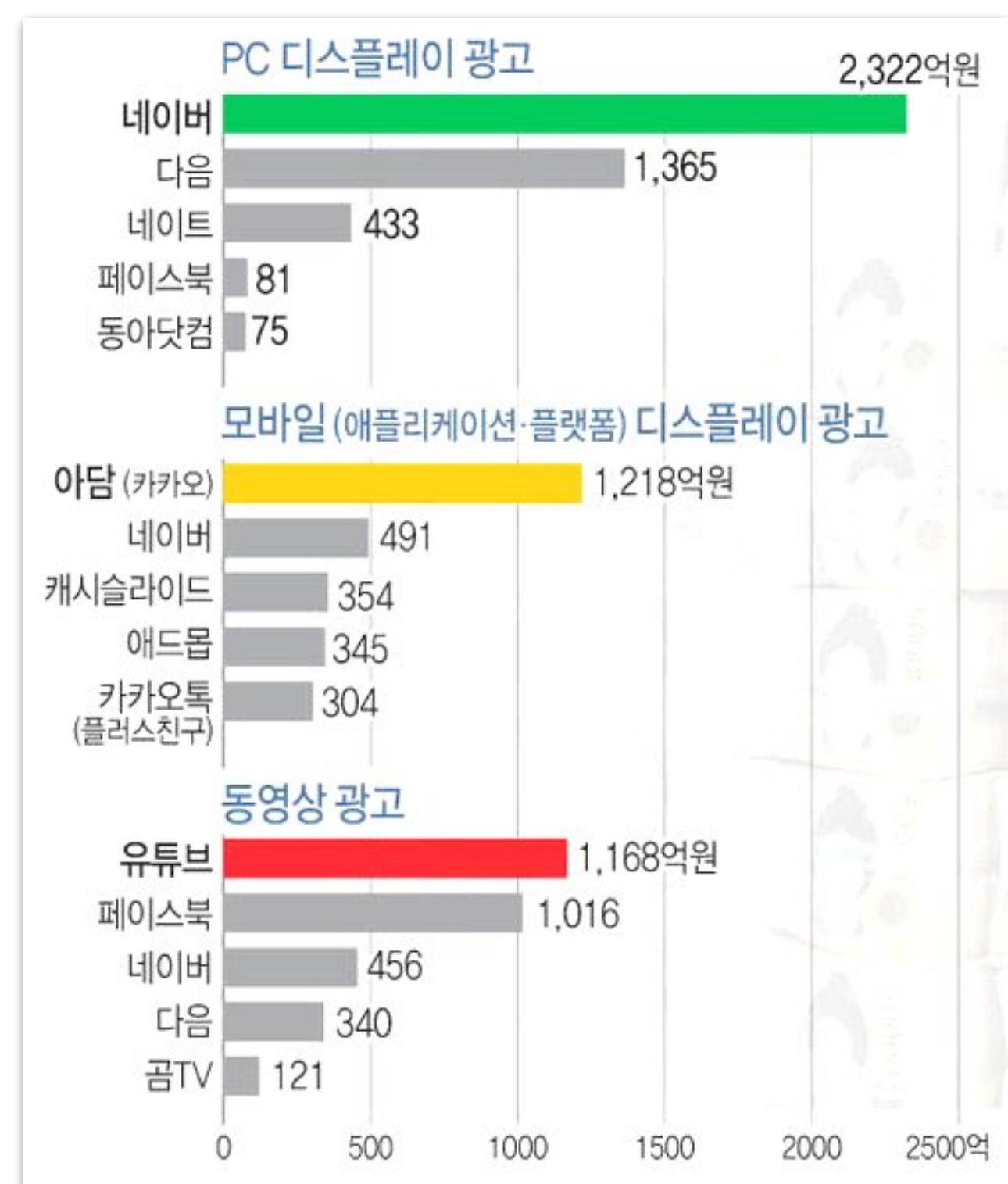
그래서 제안합니다,

2017 Top6 매체별 광고비

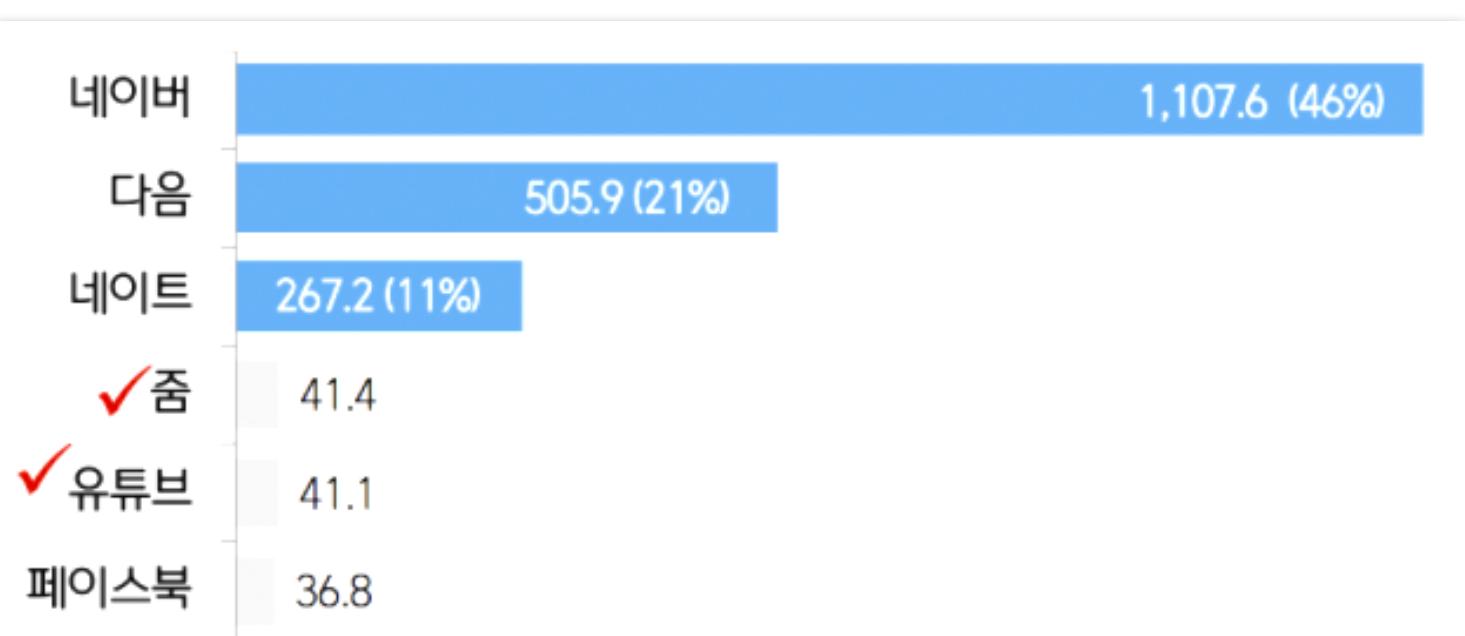


Personalized
Channel Selection

항목별 광고비 지출 상위 5개 매체



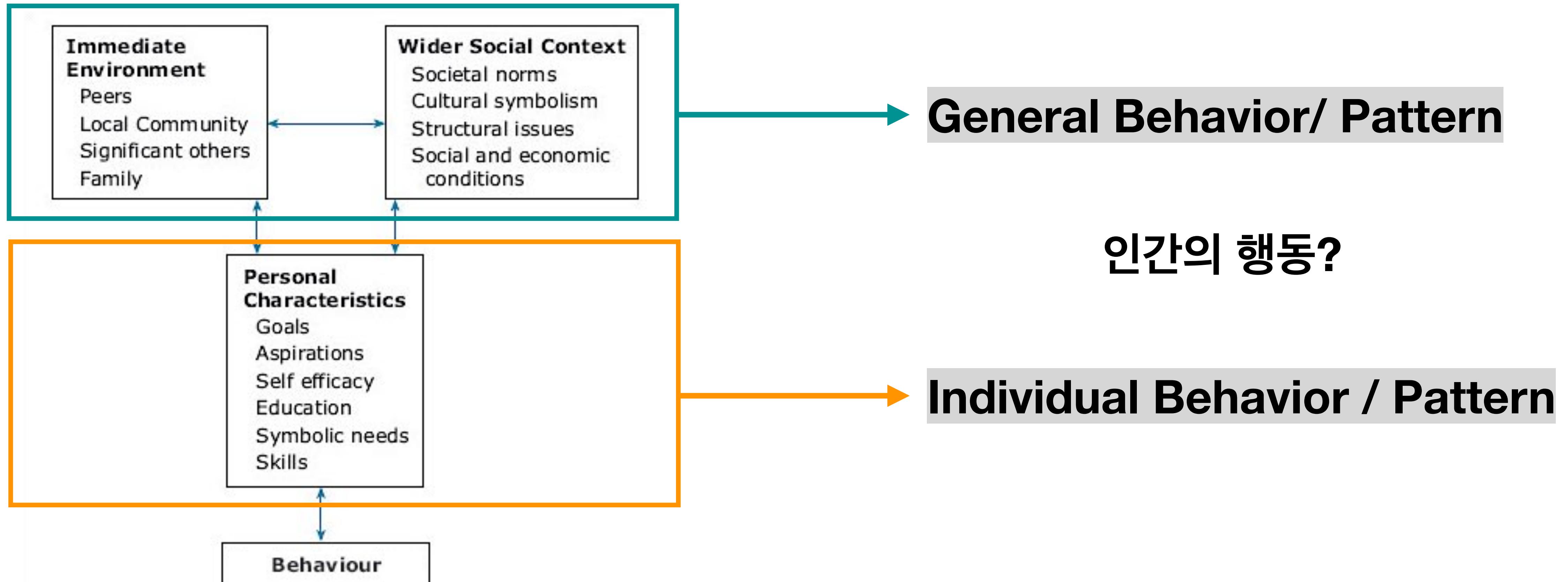
2017 Top6 매체별 광고비



Part1.

Objective

Personalized Channel Selection



Objective:

1. 스마트폰 사용 기록 속

General Pattern과 **Individual Pattern**을 확인하기

2. Individual Pattern을 활용한

Personalized Channel Recommendation 모델 만들기

Part2.

Data Analysis

Data Overview

출처: AI Hub 스마트폰 웹/앱 데이터셋

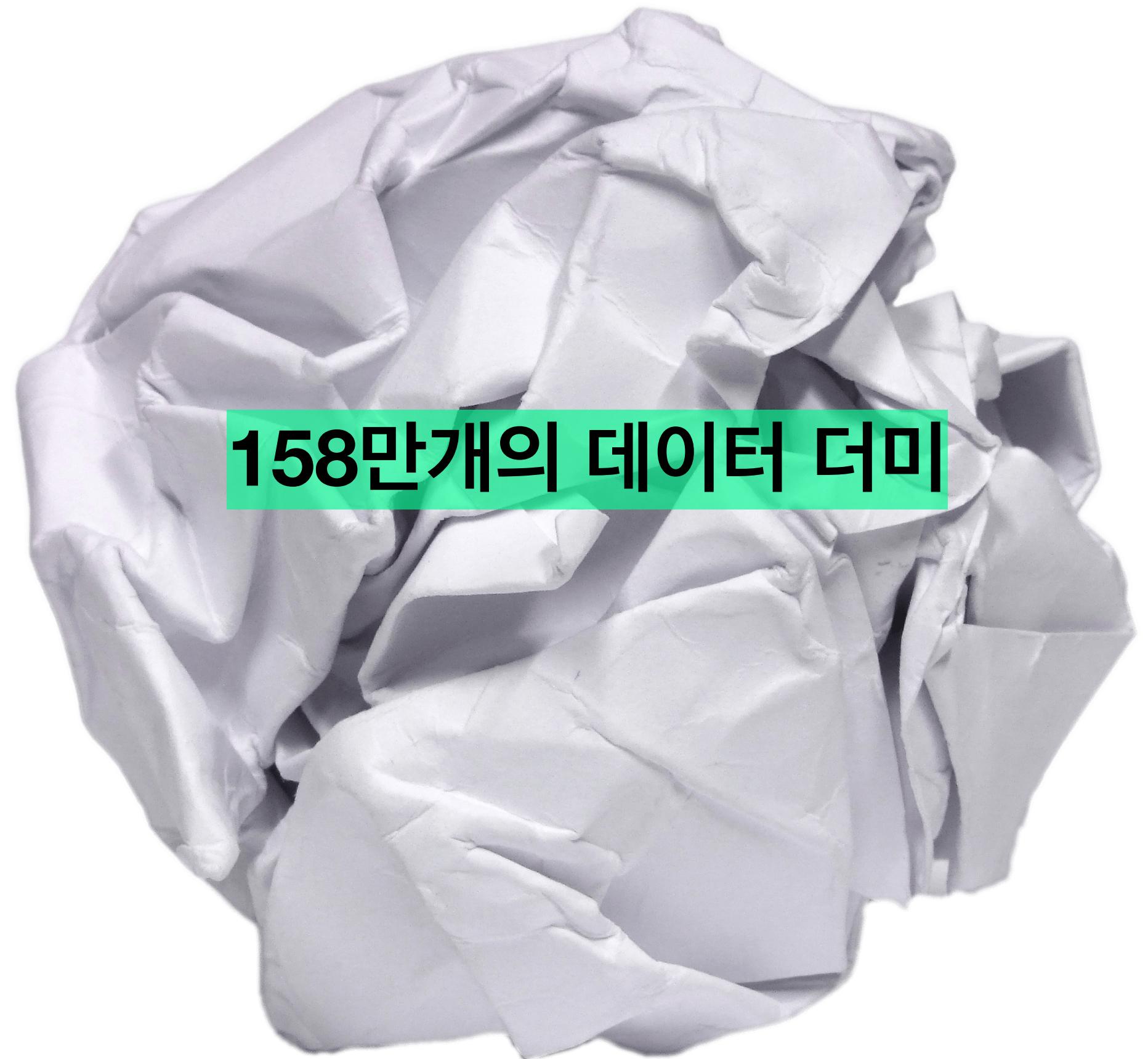
```
[{'id': 4,  
 'timestamp': 1540790989686,  
 'utcoffset': 9,  
 'subject_id': '2099efc17d719274924b8c81543c7972',  
 'source_type': 'PHONE',  
 'source_info': 'samsung-SM-G928S-7.0',  
 'package_name': 'com.nhn.android.band',  
 'name': 'BAND',  
 'is_system_app': 0,  
 'is_updated_system_app': 0,  
 'type': 'USER_INTERACTION'}]
```

Android 6.0.0 이상의 기기에서
UsageStat API를 활용해 수집한 앱 활용 내역.

31명의 피험자를 대상으로
2018.10월 - 2018.11월 중 약 3주간 수집하였으며,
총 158만개의 기록들.

JSON 형태로
사용자, 타임스탬프(시간), 사용 기기,
어플리케이션 명, 패키지명, 탑입(포그라운드-화면, 백그라운드)
등이 기록되어 있음

Preprocessing & EDA



158만개의 데이터 더미

중복 기록 제거

비정상적인 기록 제거

백그라운드 활동 등 사용자의 직접적인 활동이 아닌 기록 제거

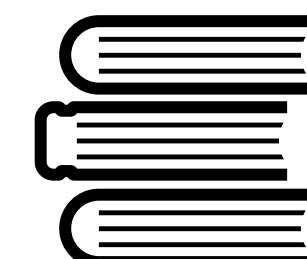
Date, Time_range: 타임스탬프 datetime으로 변경

Stay_time, Todal_sec: 어플리케이션 별 체류 시간 계산

Genre: 셀레니움 자동화 크롤링을 통해 각 어플리케이션 별 장르 칼럼 생성

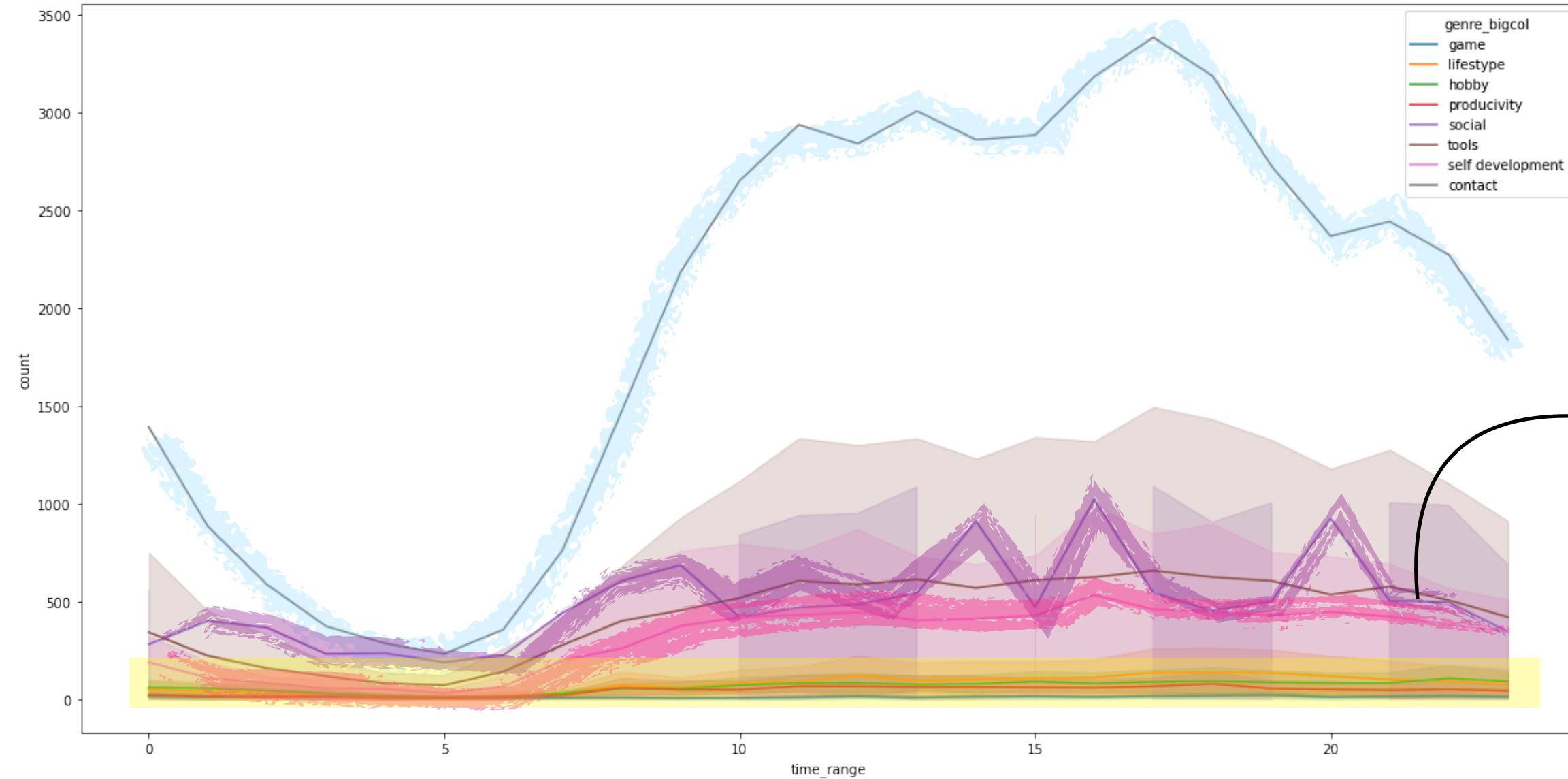
Genre_bigcol: 장르별 대장르(큰 틀) 칼럼 형성

Importance, Importance_group: 추후 설명 예정

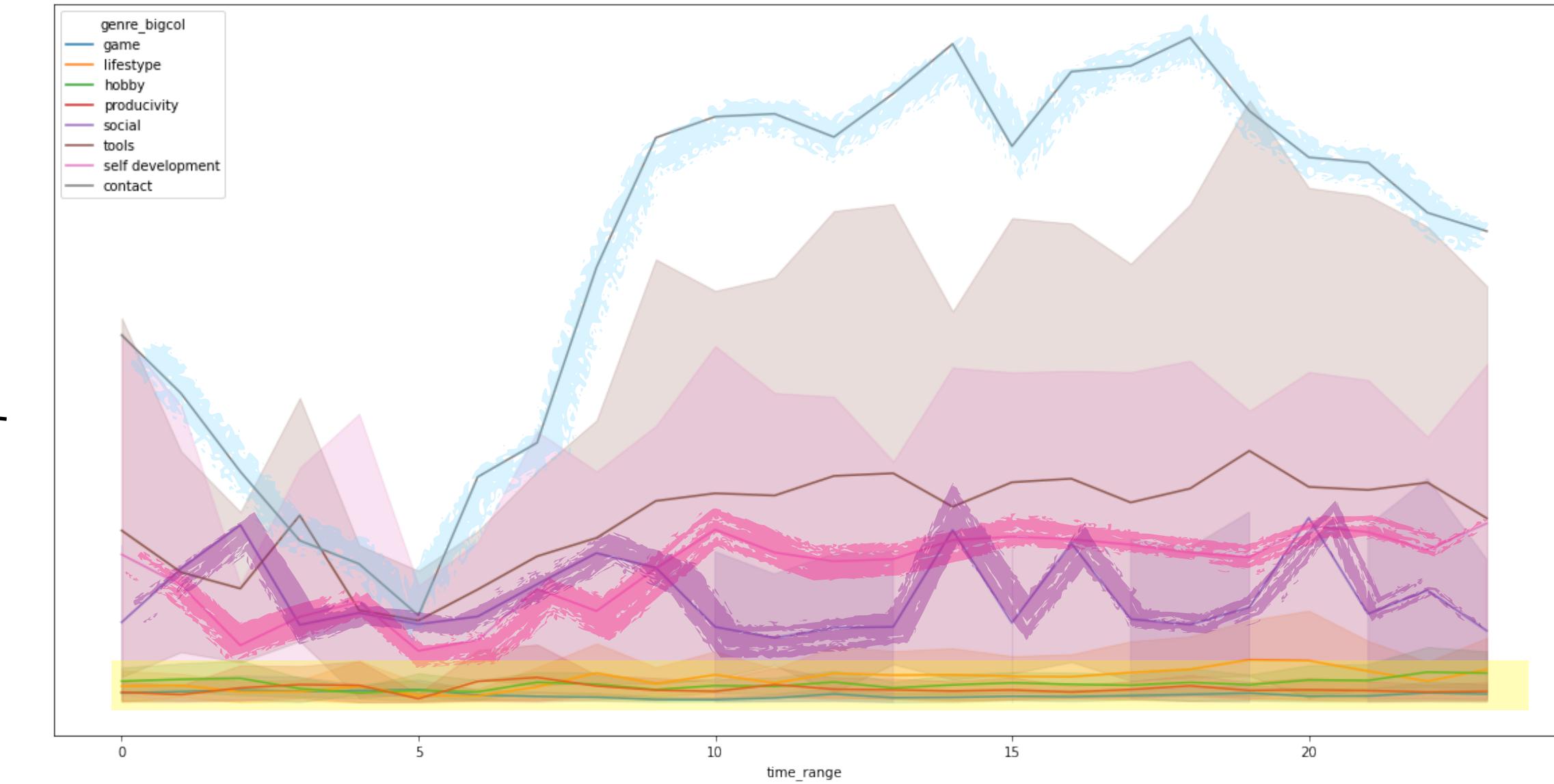


15만개의 정제된 데이터

General Usage Pattern



장르별 평균 어플리케이션 사용량 및 시간대



장르별 총 어플리케이션 사용량 및 시간대

연락용 어플리케이션: 오전 8시경부터 사용량이 급상승

소셜 어플리케이션: 들쭉날쭉한 이용량 - 사용하지 않을 때에는 0에 가깝지만, 한번 사용하면 이용량이 많아짐

자기관리 어플리케이션: 일정하면서도 꾸준한 사용량

Hobby, Lifestyle, Game 등 Entertainment: 어플 사용량은 시간대와는 관련성이 떨어짐, 시간과 관련없이 사용량이 일정

General Usage Pattern

가장 많이 사용한 어플 Top 20

genre		count	mean
	name		
Books & Reference	NAVER	3988	00:03:35
Communication	Chrome	5487	00:03:52
	메시지	3401	00:05:54
	삼성 인터넷	4187	00:04:57
	전화	2298	00:02:38
	카카오톡	26227	00:03:42
Education	ABC Platform	7630	00:05:33
Health & Fitness	Fitbit	3388	00:07:13
	캐시워크	8108	00:09:58
Music & Audio	Melon	2063	00:04:39
Personalization	Samsung Experience 흄	13527	00:06:43
	TouchWiz 흄	2981	00:08:40
	기본홈	2960	00:09:58
	시스템 UI	1528	00:05:57
Social	Between	1557	00:01:40
	Facebook	4114	00:04:18
	Instagram	3605	00:03:40
	트위터	5925	00:03:02
Tools	Samsung Galaxy Friends	1323	00:09:21
Video Players & Editors	YouTube	1884	00:08:50

가장 길게 사용한 어플 Top 20

genre		count	mean
	name		
Adventure	제5인격	139	00:09:06
Entertainment	AfreecaTV	169	00:10:14
Finance	카카오스탁	541	00:09:02
HOME	기본홈(홈&앱서랍)	621	00:15:34
Health & Fitness	Mi 피트	352	00:13:45
	캐시워크	8108	00:09:58
Lifestyle	CJ ONE	241	00:13:32
Music & Audio	음악	392	00:08:55
Personalization	기본홈	2960	00:09:58
Puzzle	위베어베어스 더퍼즐	159	00:14:04
Role Playing	검은사막 모바일	213	00:09:24
	벽猿향로	427	00:12:18
	소녀전선	314	00:09:01
Tools	AhnLab V3 Mobile Security	141	00:13:32
	Samsung Galaxy Friends	1323	00:09:21
	빅스비 흄	130	00:10:06
	시계	1206	00:09:25
	알람/시계	260	00:10:09

카카오톡, 크롬, 인스타그램 등

가장 보편적이고 기본적인 소셜, 연락, 브라우징 어플리케이션.
우리는 스마트폰의 대부분을 다음의 행위를 하기 위해 사용한다.

게임, 주식 등 개인의 ‘취미’와 관련.

각자 흥미 있는 분야에 긴 시간을 투자.

General Usage Pattern

```
[('Samsung Experience 홈', '카카오톡'), 7625),  
 ('카카오톡', 'Samsung Experience 홈'), 4335),  
 ('모아락', '캐시워크'), 2187),  
 ('허니스크린', '캐시워크'), 1927),  
 ('시스템 UI', '카카오톡'), 1877),  
 ('Samsung Experience 홈', 'NAVER'), 1849),  
 ('ABC Platform', '카카오톡'), 1736),  
 ('Samsung Experience 홈', 'Instagram'), 1641),  
 ('원더락', '캐시워크'), 1640),  
 ('Samsung Experience 홈', '삼성 인터넷'), 1570),  
 ('Samsung Experience 홈', 'Facebook'), 1272),  
 ('시스템 UI', 'Samsung Experience 홈'), 1256),  
 ('기본홈', '카카오톡'), 1170),  
 ('TouchWiz 홈', '카카오톡'), 1137),  
 ('Xperia 홈', '트위터'), 1129),  
 ('Xperia 홈', '카카오톡'), 1098),  
 ('Fitbit', '카카오톡'), 1088),  
 ('Samsung Experience 홈', 'ABC Platform'), 1043),  
 ('연락처', '전화'), 834),  
 ('시스템 UI', 'Chrome'), 828)]
```

핸드폰을 열었을 때 가장 많이 하는 행위는?

1. '카카오톡'

가장 빈번한 이동: 기본 홈 <-> 카카오톡
핸드폰을 열자마자 가장 빈번하게 사용하는 어플리케이션 '카카오톡'

이외에도 허니스크린, 모어락, 시스템 UI등 다른 종류의 기본홈이나
페이스북, 웹 브라우저 등 다른 어플에서
카카오톡으로 이동하는 현상도 빈번하게 나타남

다른 어플 사용 중 연락을 위해 카카오톡으로 이동한 것으로 보임.

2. '브라우저'

기본 홈~네이버로의 이동도 빈번.
네이버 이외에도 크롬, 인터넷 등 다른 웹서핑 브라우저로의 이동 또한 빈번.

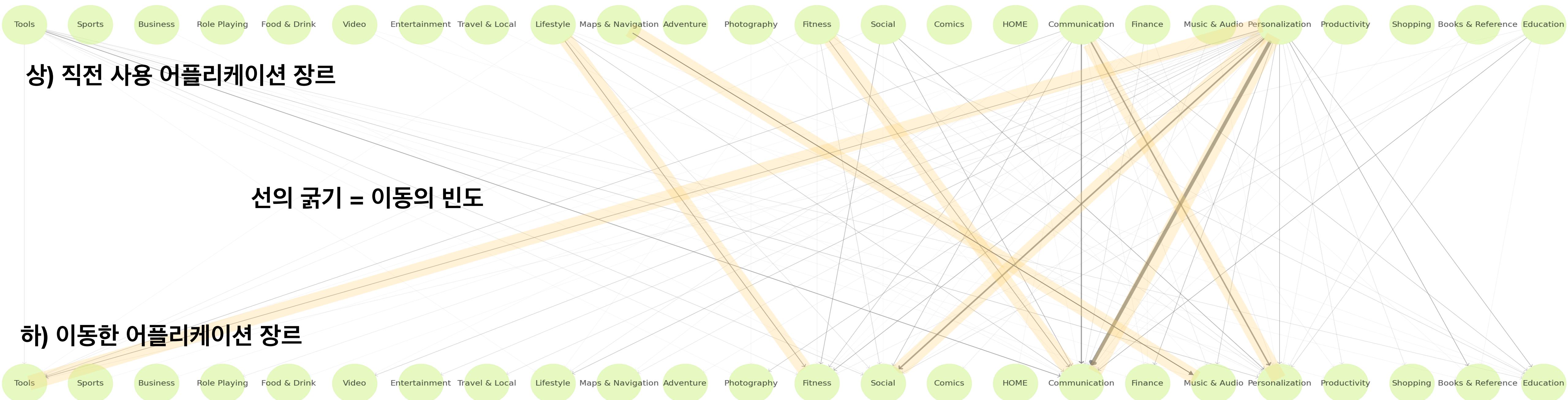
3. '인스타그램'

대세 SNS답게
카카오톡, 네이버 다음으로 핸드폰을 열었을 때 가장 빈번히 클릭하는 어플.

General Usage Pattern

Referrer: 이전에 사용한 어플

어플 간 이동에서, 이전에 사용한 어플이 이후 사용할 어플에 영향을 미칠까?



Personalization(기본홈) -> 커뮤니케이션

커뮤니케이션 -> 커뮤니케이션

지도 -> 음악

등의 어플 간 이동 패턴

Individual Usage Pattern

각자의 Individual한 행동 패턴은 어떻게 구분할 수 있을까?



기준 1.

나의 이 취향이 나만의 것인가?

취향 지수 = (개인의 이용에서 특정 어플 이용이 차지하는 비율)
/ (전체 이용에서 특정 어플 이용이 차지하는 비율)

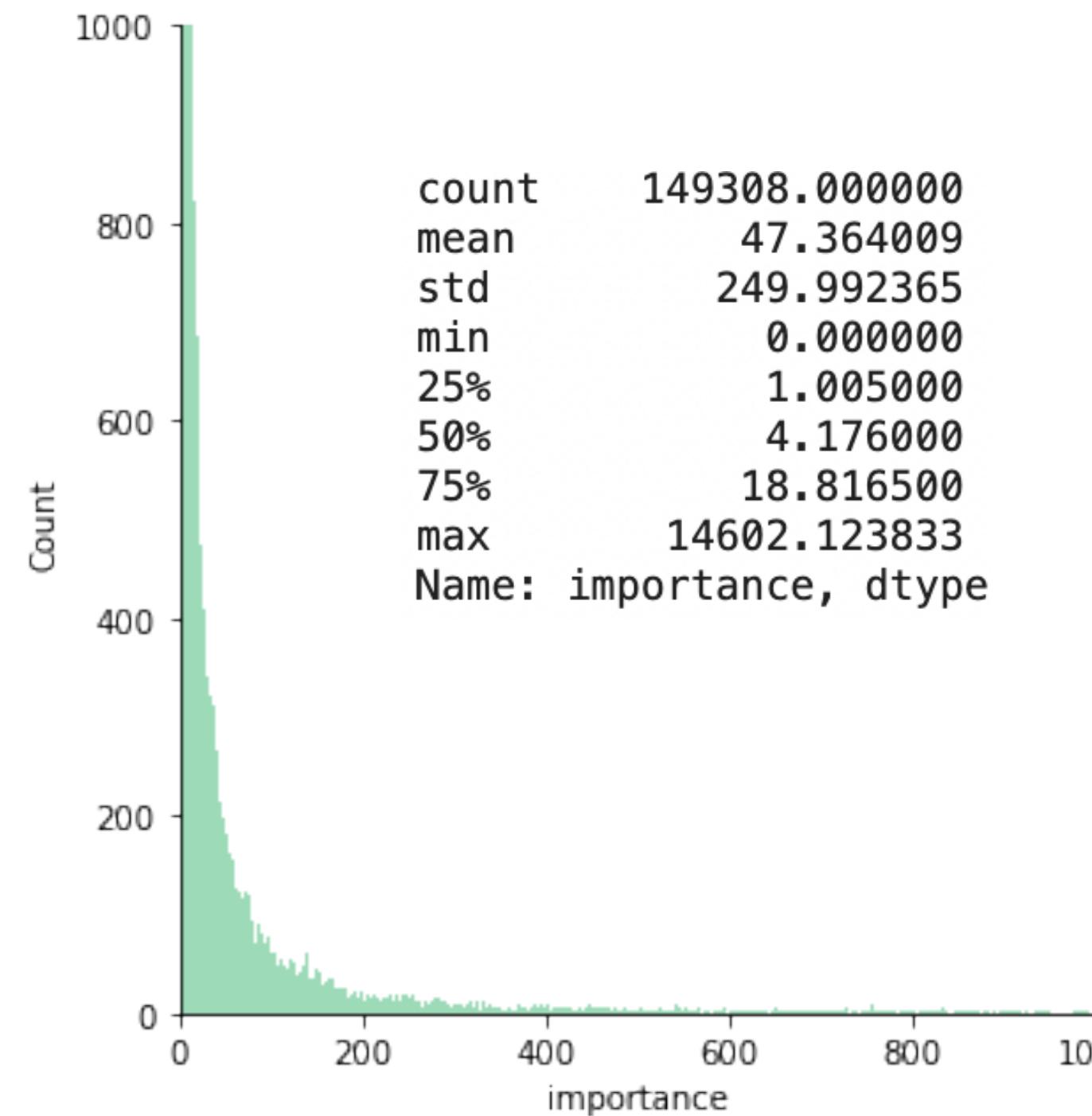
기준 2.

해당 행위에 충분한 시간을 투자하는가?

총 사용시간 (stay_time)

Individual Importance level

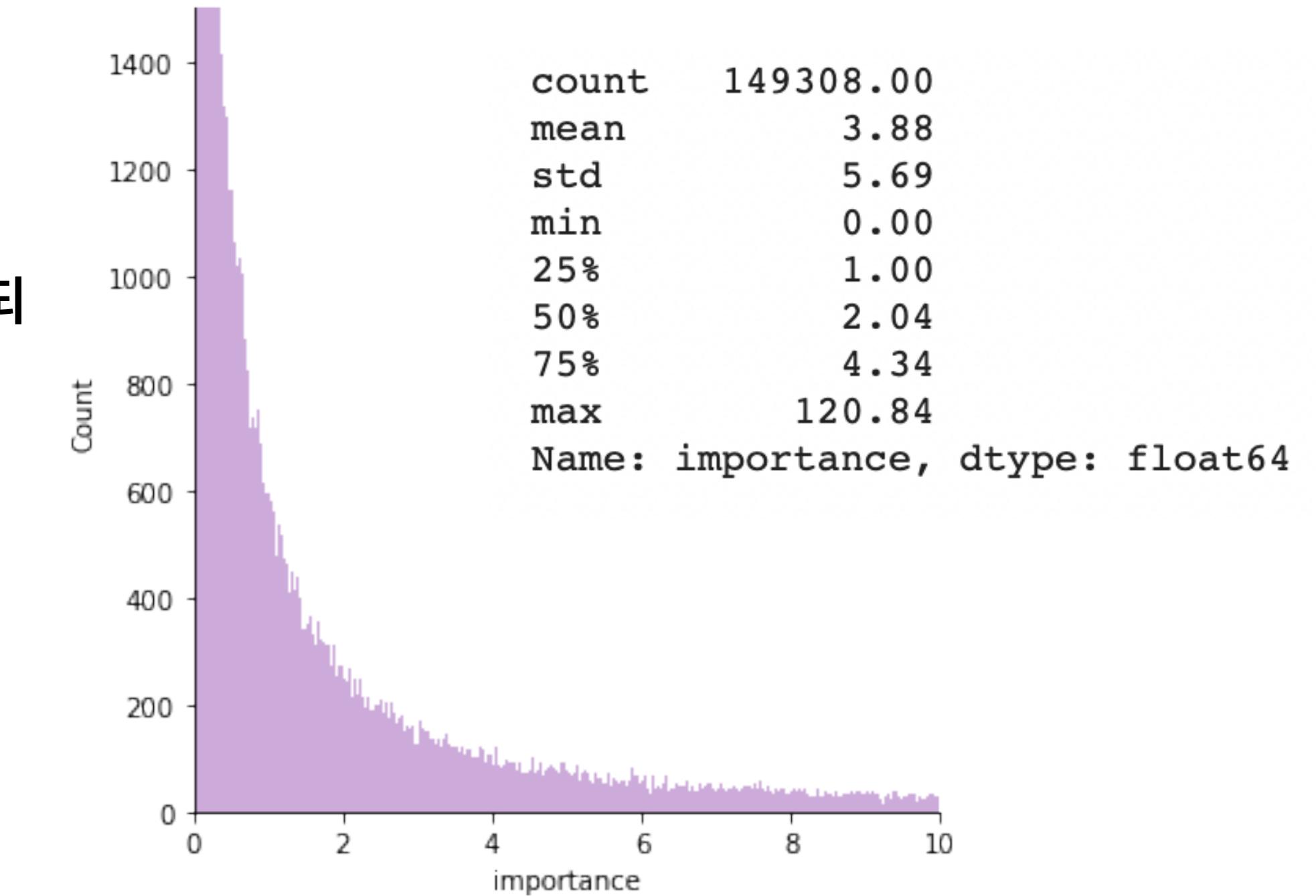
특징 지수



기존 취향 지수의 분포 그래프
왼쪽(0)으로 치우친 편차를 보이나,
최대값이 14000에 달할 정도로 지수 간 편차가 큼

Sqrt
→

루트를 써워 편차를 조절하되
데이터의 형태 및 분포는
그대로 유지하고자 함.



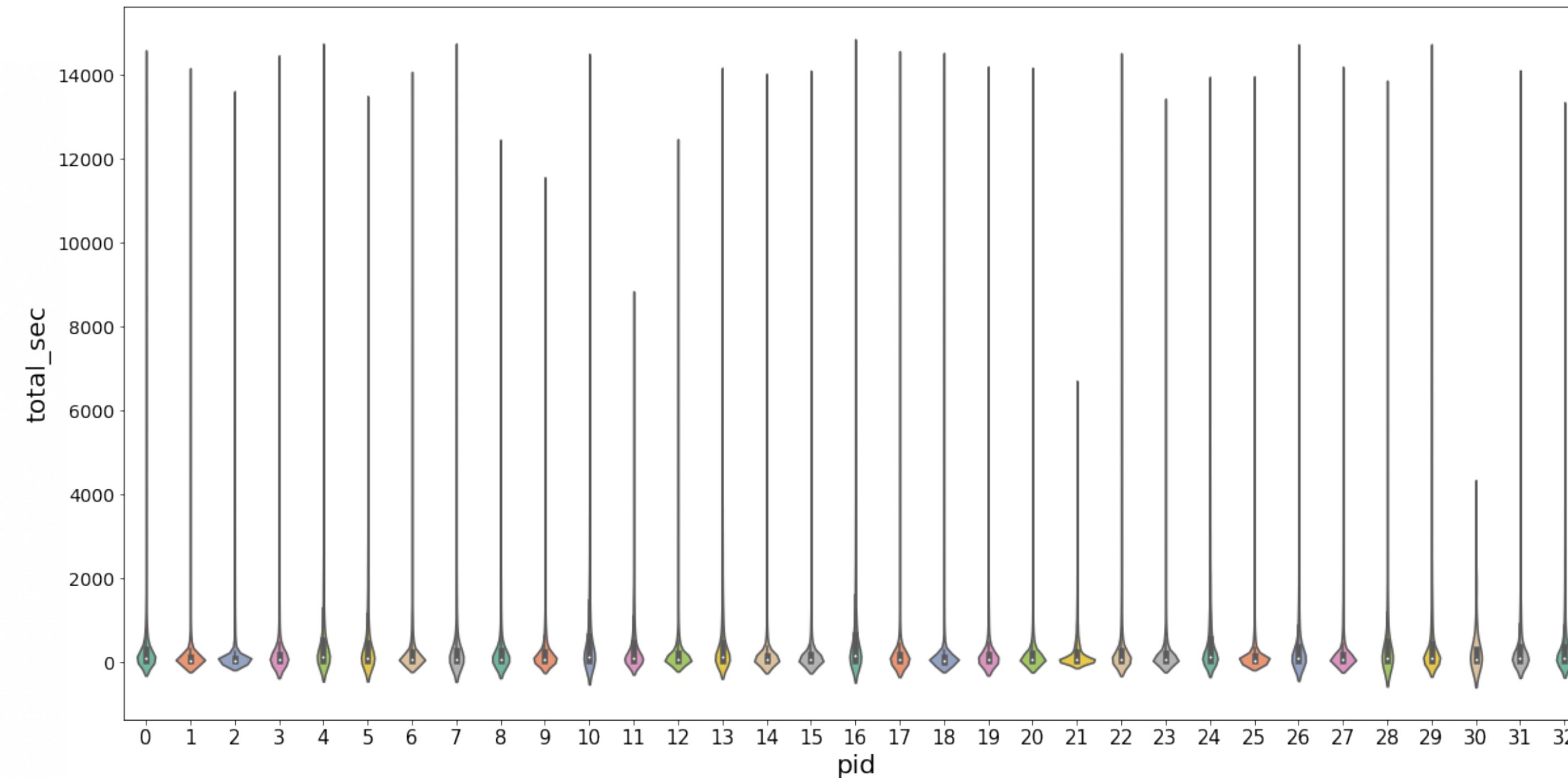
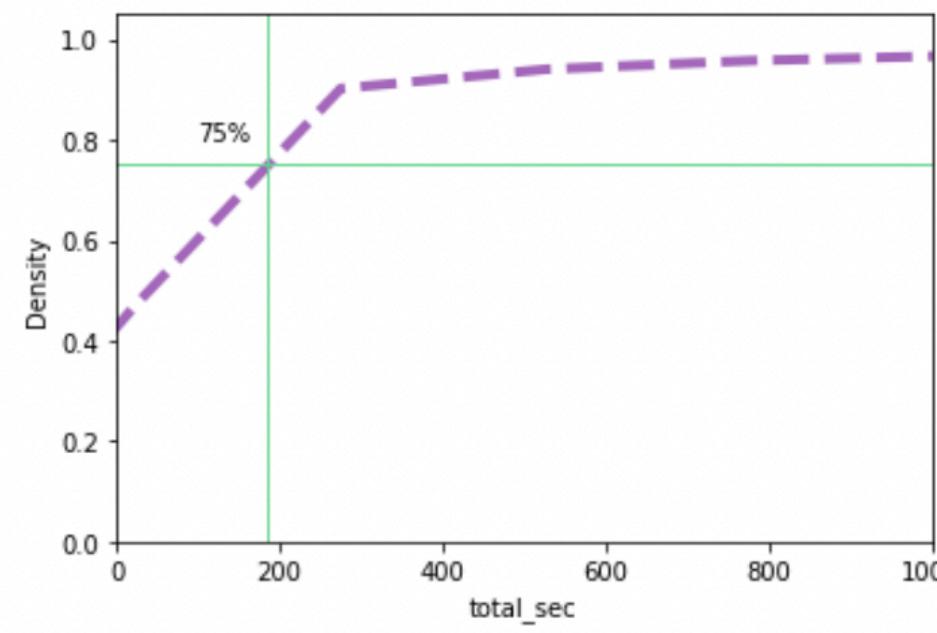
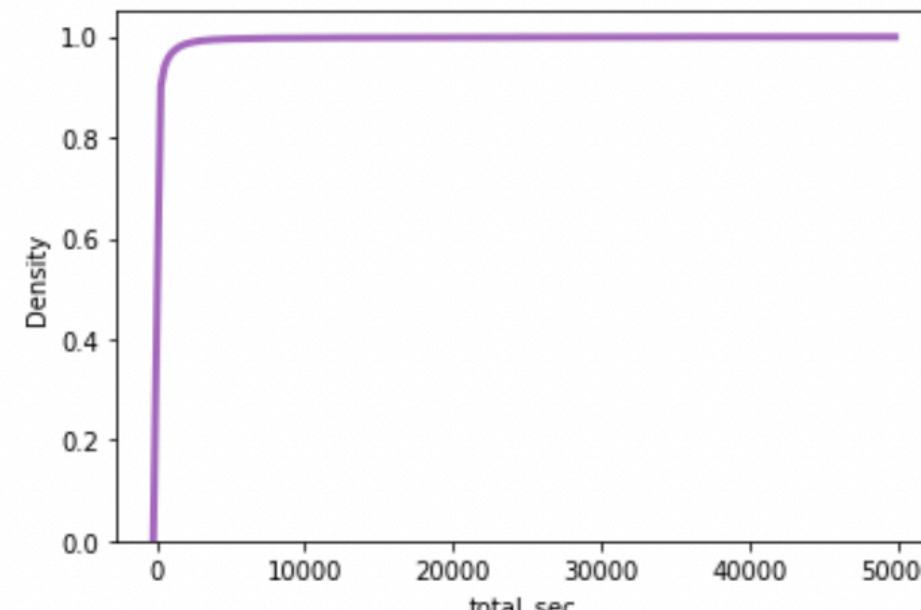
루트 변환된 취향지수 분포 그래프
취향 지수 분포에서 75% 수준을 기준으로 삼아
취향 지수 4 이상의 행위를 개인의 특징을 반영하는 특이값으로 봄

Individual Smartphone Usage Time

사용 시간

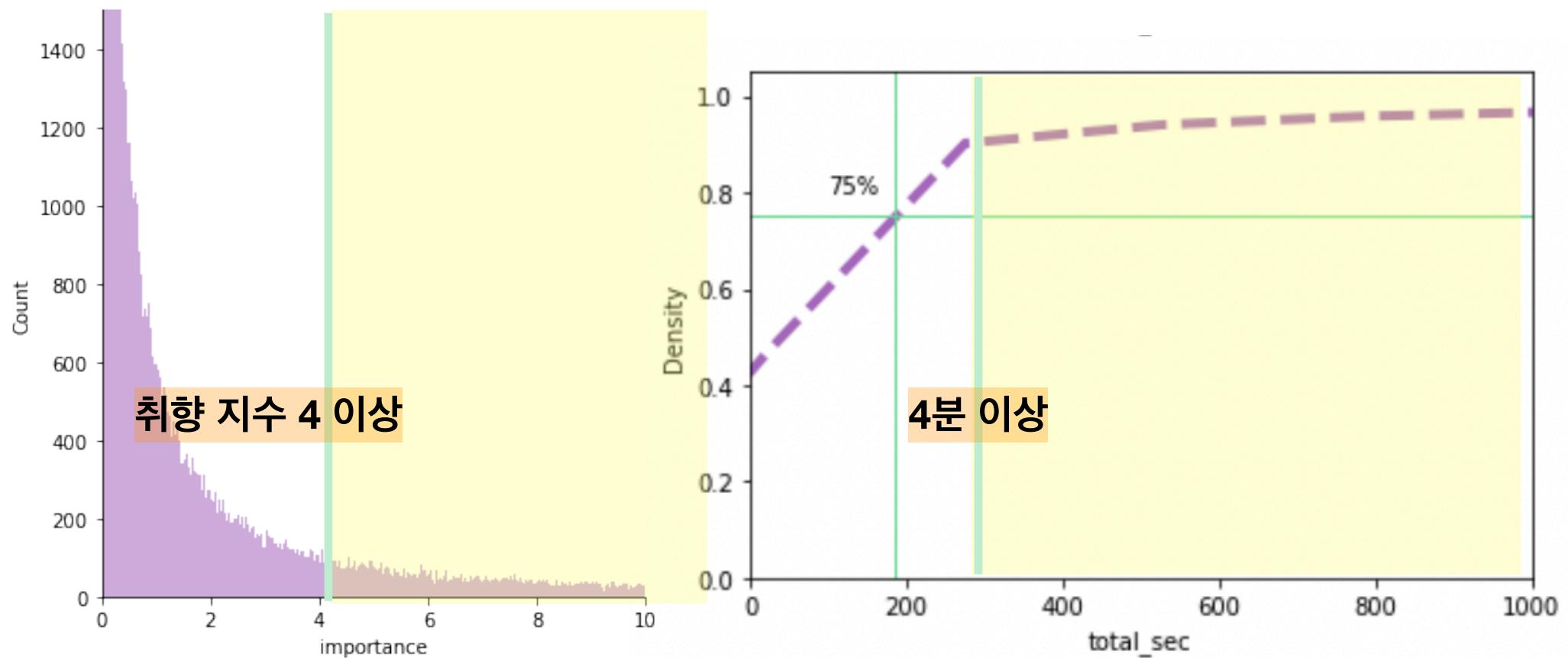
```
np.array(how_many_wakeup).mean()
```

215.52433197827966



우리는 스마트폰을 짧게, 빈번히 이용한다.

전체 스마트폰 이용 행위 시간에서도 75% 수준인 '4분 이상 사용행위'를 특이값으로 봄



취향 지수, 사용 시간의 75% 지점을 Threshold로 설정,
각 지점은 취향 지수 4, 사용 시간 4분
이를 바탕으로 행동을 다음 4가지로 구분

취향 지수 4 이하, 사용 시간 4분 이하 -> 행동 1

취향 지수 4 이상, 사용 시간 4분 이하 -> 행동 2

취향 지수 4 이하, 사용 시간 4분 이상-> 행동 3

취향 지수 4 이상, 사용 시간 4분 이상 -> 행동 4

남도 ← → 나만?

조금

오래

행동 2

남도 하고,
조금 함

행동 1

나만 하는데,
조금 함

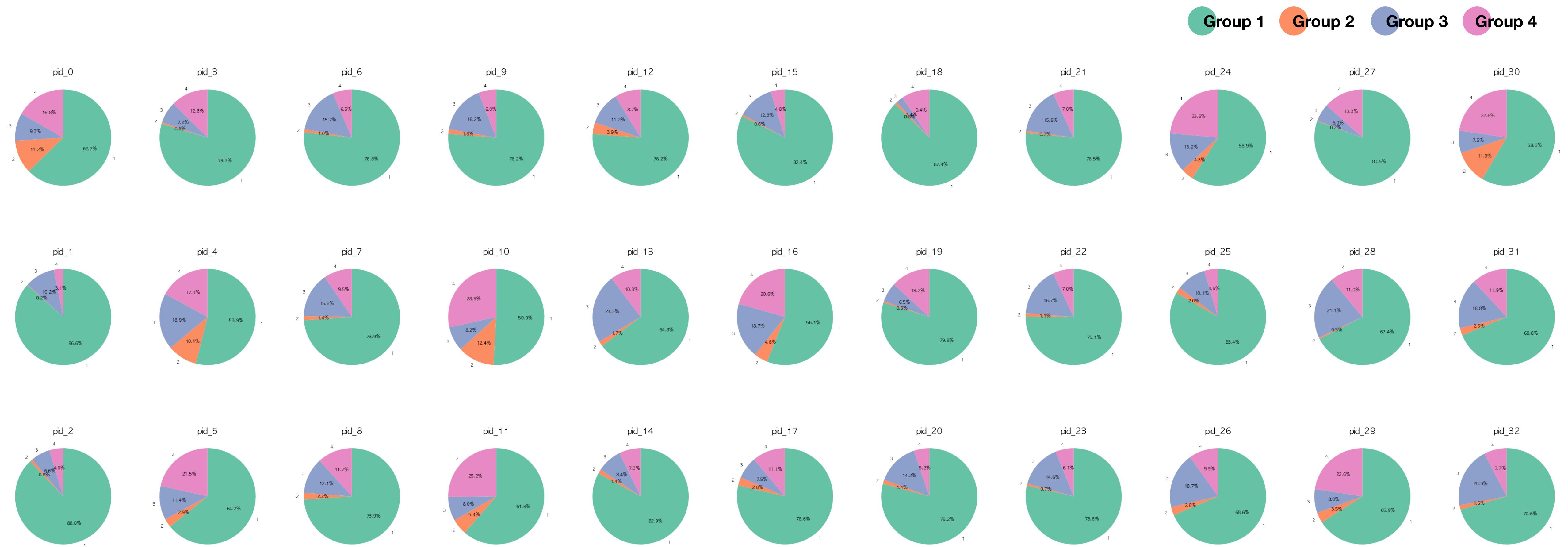
행동 3

남도 하지만,
오래 함

행동 4

나만 하는데,
오래 함

Individual Usage Pattern



행동 그룹별 개인의 스마트폰 행위 구성 비율

그룹 (1) (개인만의 특성도 아니고, 이용 시간도 짧은 행위)가 70%를 차지

다만 개인만의 특성을 반영하는 중요도(4)의 행위가 그 다음으로 빈번하게 나타남

Individual Smartphone Usage Plot

X = Application Name, Genre, Big Category of Genre

Y = Time Range

Size = Amount of Time
Color = Importance

Group 1

Group 2

Group 3

Group 4

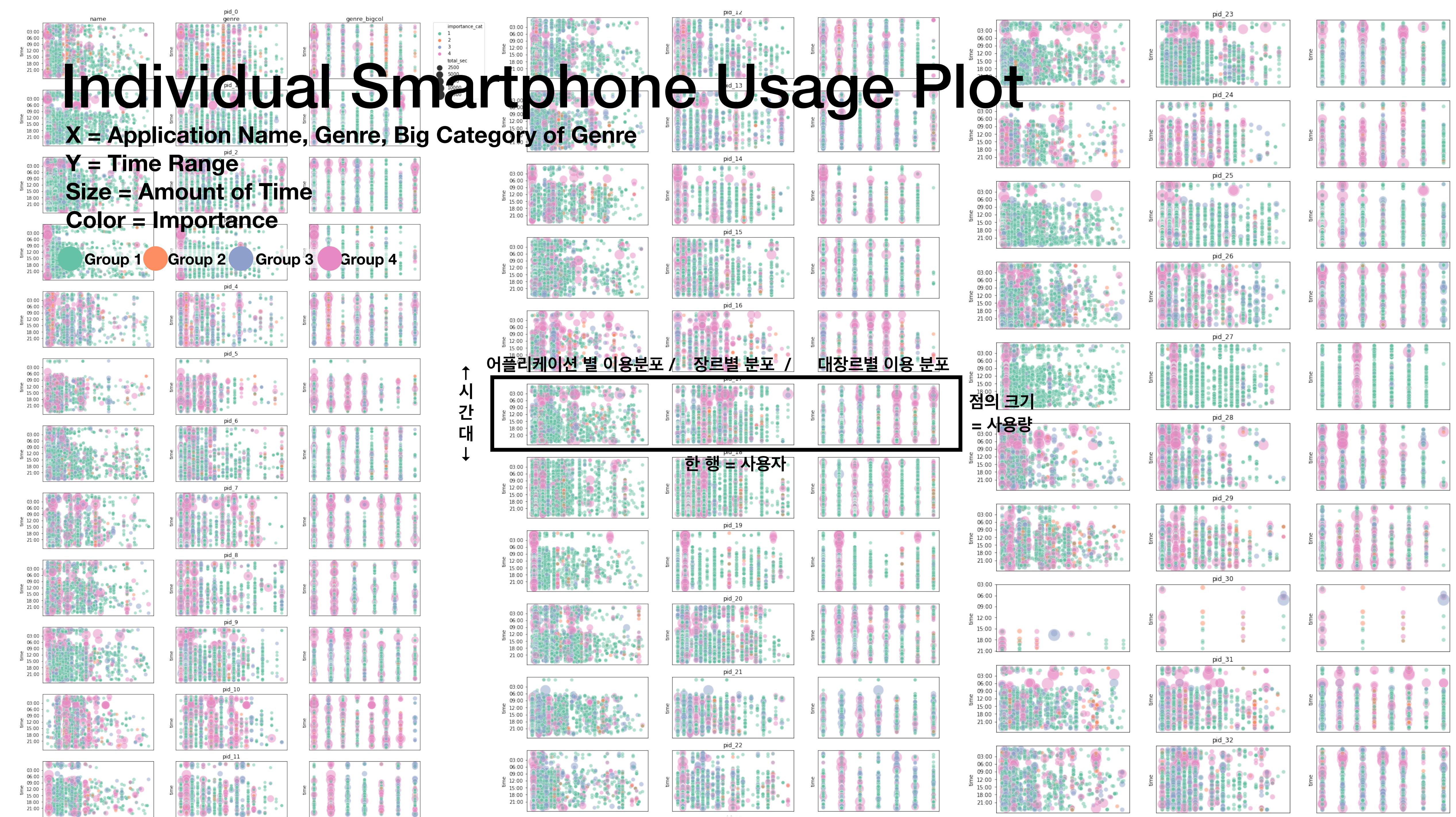
↑ 시간 대 ↓

어플리케이션 별 이용분포 / 장르별 분포 / 대장르별 이용 분포

점의 크기

= 사용량

한 행 = 사용자



Individual Smartphone Usage Plot

X = Application Name, Genre, Big Category of Genre

Y = Time Range

Size = Amount of Time

Color = Importance

Group 1

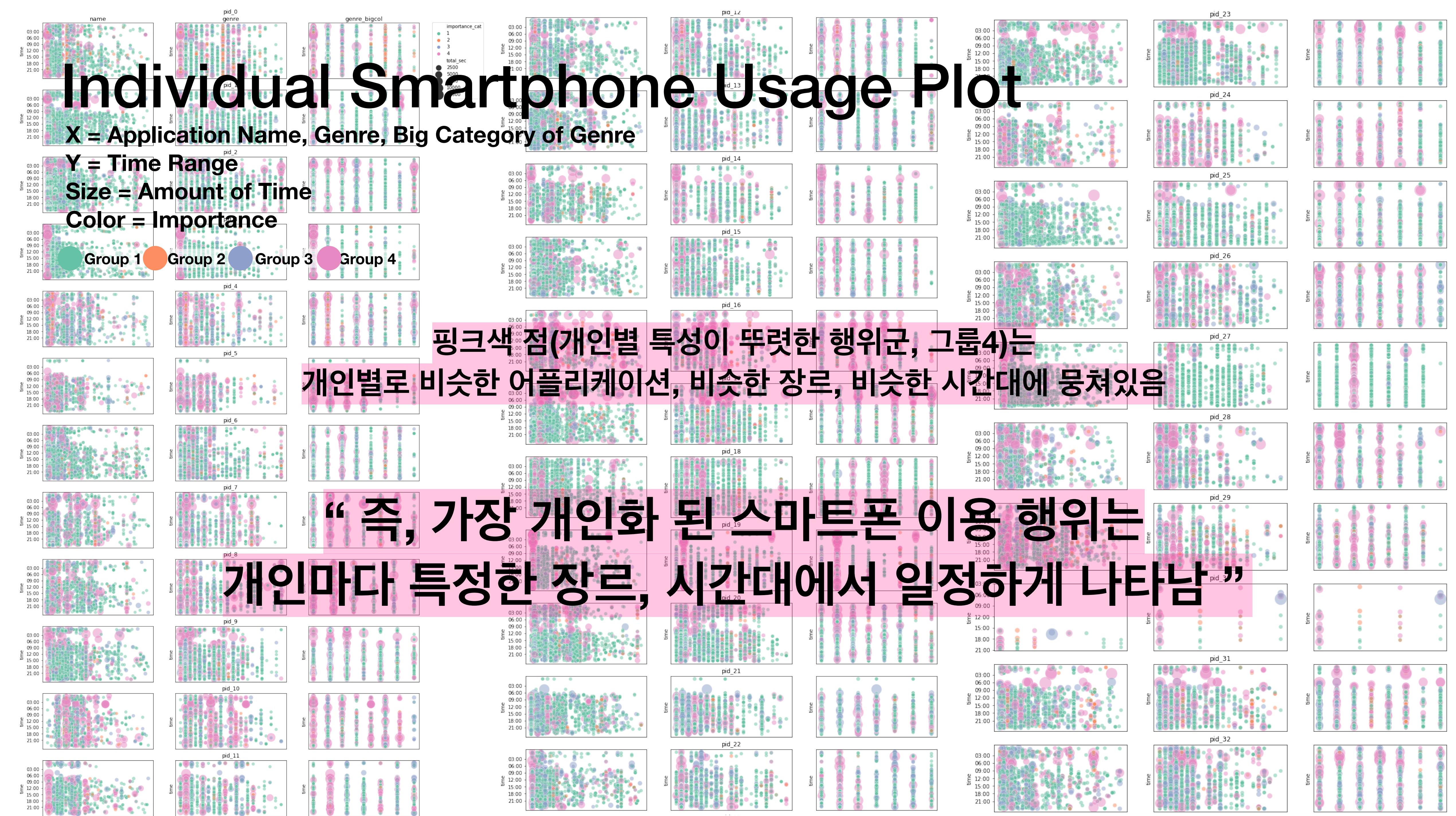
Group 2

Group 3

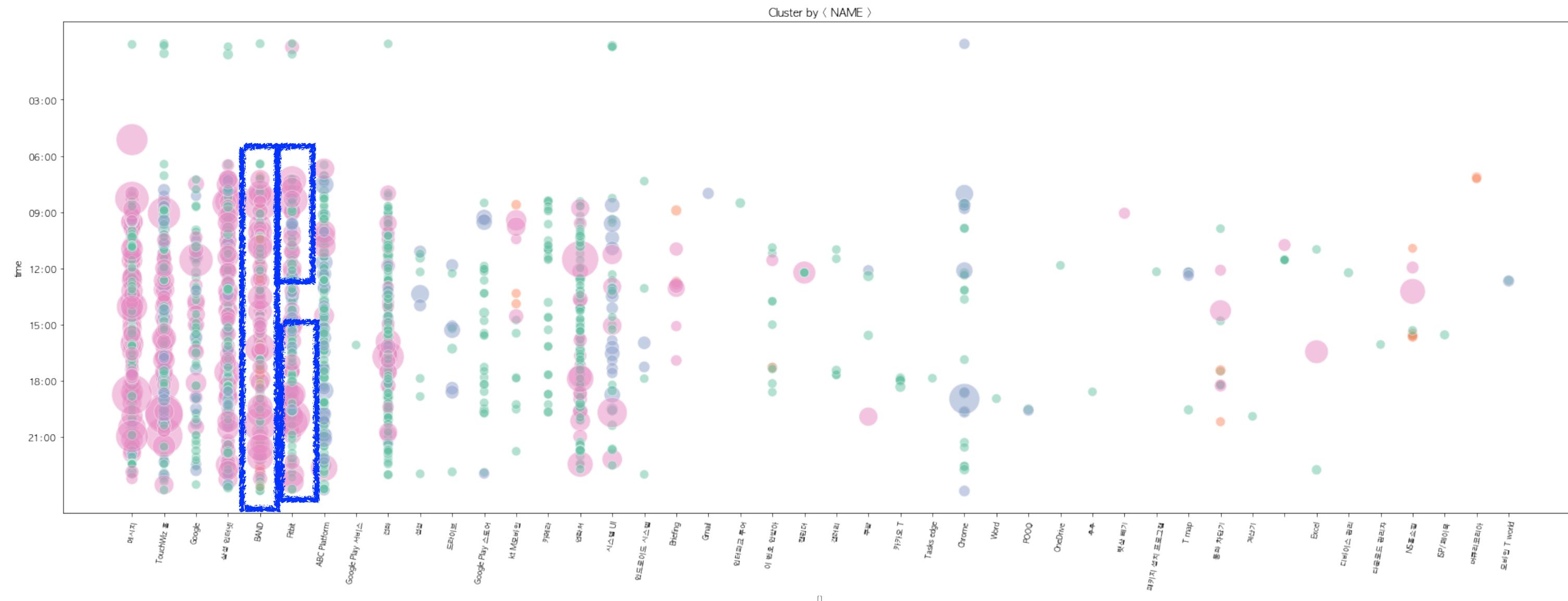
Group 4

핑크색 점(개인별 특성이 뚜렷한 행위군, 그룹4)은 개인별로 비슷한 어플리케이션, 비슷한 장르, 비슷한 시간대에 뭉쳐있음

“즉, 가장 개인화 된 스마트폰 이용 행위는 개인마다 특정한 장르, 시간대에서 일정하게 나타남”

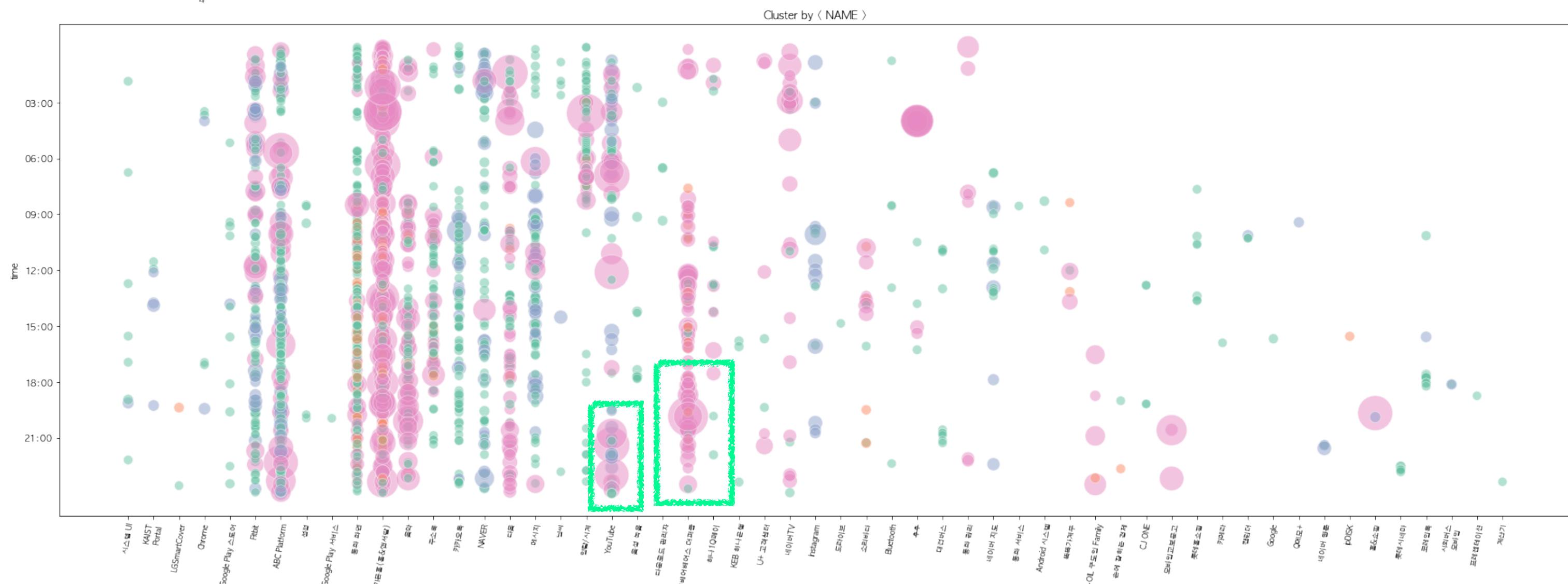


Individual Usage by Personal Importance



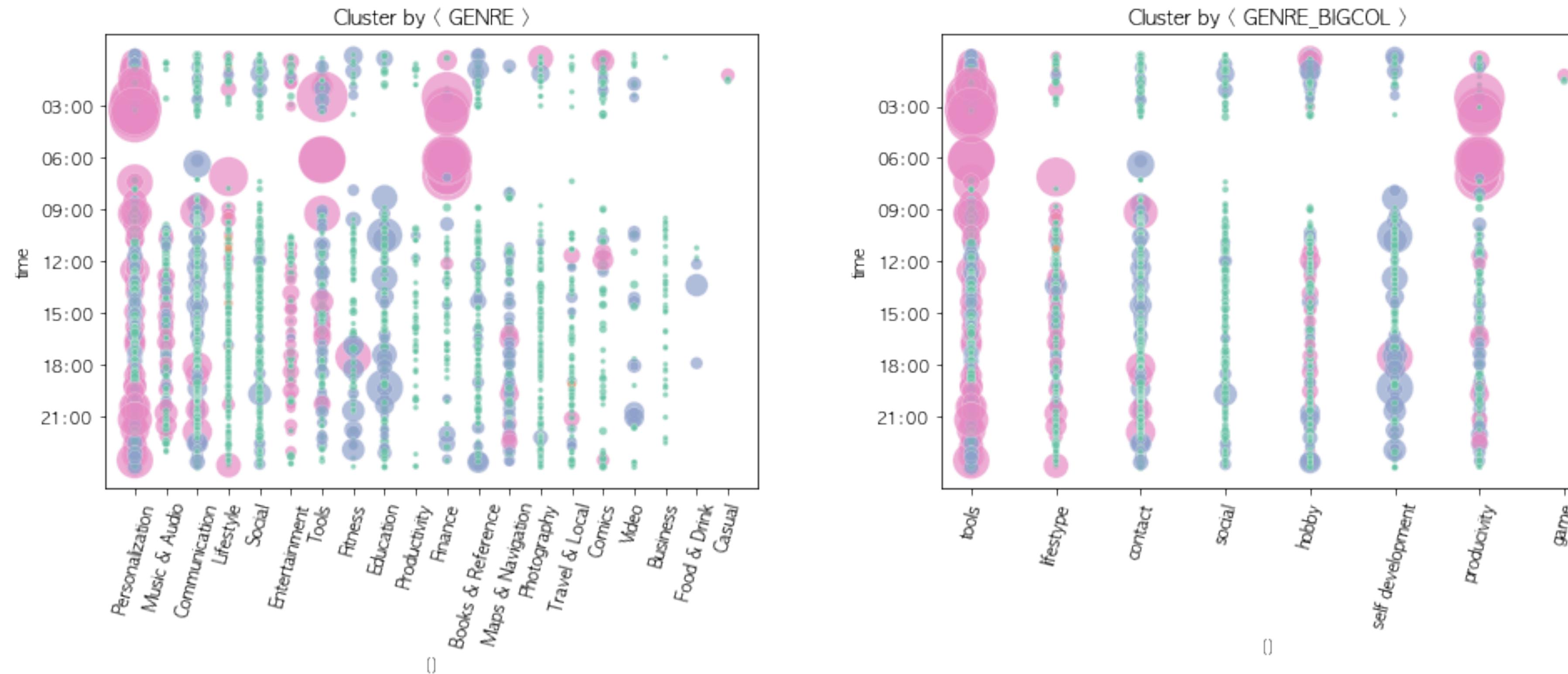
**저녁 7시-8시 사이에 유튜브를 자주 보고,
특히 저녁에 위비어베어스 더퍼즐을 즐겨하는
ID 10번 씨**

아침 8시경, 저녁 9시경
샤오미 핏비트로 운동하고,
소셜 어플리케이션 ‘BAND’를 많이 사용하는
id 1번씨



Individual Usage by Personal Importance

pid = 10



가장 개인화된 행동이 나타나는 행동 그룹 4는
개개인의 취향대로 특정한 장르, 특정한 시간대에서 확실한 군집성을 보여줌

Part3.

Creating a ML/DL Model

Artificial Intelligence Model

사용자 기준 기록

사용 패턴 분석

미래 사용할 어플리케이션 채널 예측



개인화된 채널 선택을 통해 광고비 효율 개선

Machine Learning Model

LGBM 분류기

각 소비자별 선호 + 어플리케이션 사용 패턴을 바탕으로

각 소비자가 다음에 사용할 어플리케이션 ‘대장르’를 예측하도록 함

대장르는 총 8가지,

Baseline Model Accuracy = 0.125

	precision	recall	f1-score	support
contact	0.33	0.42	0.37	4063
game	0.28	0.30	0.29	137
hobby	0.19	0.18	0.19	1160
lifestyle	0.17	0.02	0.04	369
productivity	0.13	0.26	0.17	1380
self development	0.31	0.59	0.41	2414
social	0.00	0.00	0.00	1750
tools	0.54	0.18	0.28	3656
accuracy			0.30	14929
macro avg	0.25	0.24	0.22	14929
weighted avg	0.31	0.30	0.27	14929

평균 F1스코어가 0.3, 부족한 성능

-> 머신 러닝 대신 딥러닝 모델을 시도

Deep Learning Model

각 소비자별 선호 + 어플리케이션 사용 패턴을 바탕으로

각 소비자가 다음에 사용할 '어플리케이션 명'를 예측하도록 함

어플리케이션은 모두 693개.

Baseline Model Accuracy: $1 / 693 * 100 = 0.001$

TRAIN:

타겟- 사용 어플리케이션명

특성- 개인ID, 사용 어플리케이션명, 어플리케이션 장르, 행위 그룹(1,2,3,4), 시간,
사용 시간대, Referrer(현재 어플리케이션으로 이동하기 전에 사용한) 어플리케이션명, Referrer 장르, Referrer 패키지

MODEL:

훈련 데이터를 벡터화, 이를 임베딩 층에 가중치로 주입하고, LSTM 사용.

TEST:

미래 사용할 어플리케이션을 예측할 때에는 어플리케이션을 알지 못하므로 그 행위가 어떤 그룹에 속할지를 알 수는 없음.
이에, 관심도를 주지 않은 상태로 예측을 시켜 봄.

Layer (type)	Output Shape	Param #
embedding_13 (Embedding)	(None, 10, 238)	28424816
dense_63 (Dense)	(None, 10, 256)	61184
lstm_11 (LSTM)	(None, 256)	525312
dense_64 (Dense)	(None, 128)	32896
dense_65 (Dense)	(None, 64)	8256
flatten_6 (Flatten)	(None, 64)	0
dense_66 (Dense)	(None, 696)	45240
<hr/>		
Total params: 29,097,704		
Trainable params: 29,097,704		
Non-trainable params: 0		

Deep Learning Model

	precision	recall	f1-score	support
0	0.88	0.97	0.92	271
1	0.51	0.59	0.55	823
2	0.99	0.99	0.99	2650
3	0.59	0.56	0.58	246
4	0.44	0.87	0.58	75
5	0.27	0.09	0.13	146
6	0.80	0.75	0.78	662
7	0.71	0.04	0.08	115
8	0.95	0.93	0.94	1535
9	0.91	0.84	0.87	5334
10	0.00	0.00	0.00	8
11	0.83	1.00	0.91	5
12	0.00	0.00	0.00	18
13	0.24	0.26	0.25	66
14	0.92	0.99	0.95	72
15	0.95	0.97	0.96	389
16	0.69	0.31	0.42	36
17	0.87	0.78	0.82	100
18	0.92	0.87	0.89	324
19	0.30	0.52	0.38	23
20	0.15	0.13	0.14	15
21	0.00	0.00	0.00	3
22	0.00	0.00	0.00	7
23	0.70	0.44	0.54	449
24	0.25	0.80	0.38	10
25	0.94	0.95	0.95	439
26	0.20	0.50	0.29	2
27	0.50	0.12	0.20	57
28	0.00	0.00	0.00	2
29	0.00	0.00	0.00	3
30	0.34	0.41	0.37	125
31	1.00	0.14	0.25	7

	accuracy	macro avg	weighted avg	
0	0.75	0.26	0.28	29859
1	0.25	0.26	0.28	29859
2	0.75	0.77	0.75	29859

(단순히 평균을 계산하는 것이 아닌,
각 class에 해당하는 data의 개수에 가중치를 두어 평균을 구한 것)

Weighted Average 0.75

타겟의 High Cardinality 특성 상
수가 적은 Records들에 대해서는 정확도가 떨어지지만,
어느정도 충분한 기록이 있는 타겟 Class에 대해서는 높은 예측 정확도를 보임.